

Econometria

Heterocedasticidade



Heterocedasticidade

- Uma hipótese importante do modelo clássico de regressão linear é que a variância de cada termo de erro u_i , condicional aos valores selecionados das variáveis explicativas, é constante.
- Essa é a hipótese de **homocedasticidade**; "HOMO": igual e "CEDASTICIDADE": espalhamento.
- $Var(u_i) = \sigma^2, \quad i = 1, \dots, n$
- Considere o problema de explicar despesas com alimentação (y) como função da renda familiar (x).
- **PERGUNTA:** Você acha que seria mais fácil estimar a despesa com alimentação em uma família de baixa renda ou em uma família de renda alta? Ou não teria diferença?

Heterocedasticidade

- As famílias de renda baixa, em geral, não têm a opção de gostos extravagantes em sua alimentação e por isso têm menos escolhas.
- Famílias com rendas altas podem ter gosto simples ou extravagantes quanto à alimentação. Podem jantar, por exemplo, camarão ou omelete.
- Assim, a renda é menos importante como variável explicativa para despesa com alimentação de uma família de renda alta.
- Este tipo de problema pode ser resolvido por um modelo estatístico que exiba heterocedasticidade.

Exemplo 1: Relação entre renda familiar e despesas com alimentação (DadosExemplo1Aula2.txt)

- a) Faça o gráfico de y versus x e apresente a reta estimada.
- b) É possível verificar algo no gráfico à medida que a renda cresce?

Exemplo 1

Heterocedasticidade

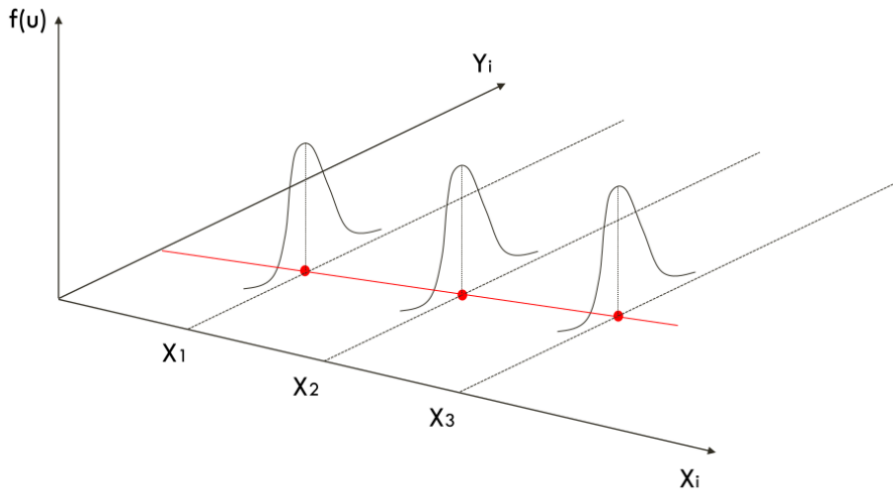
- Notem que, à medida que a renda cresce os pontos (x_i, y_i) referentes aos dados observados tendem a afastar-se da função média estimada.
- Ou seja, os resíduos de mínimos quadrados aumentam, em valor absoluto, à medida que a renda cresce.
- Como os resíduos de mínimos quadrados observáveis são estimativas aproximadas dos erros não observáveis, a figura também sugere que os erros também aumentam à medida que a renda cresce.
- Confirmando assim a hipótese de que a equação de despesa média é melhor para explicar a despesa com alimentação de uma família de baixa renda do que de uma família de renda alta.

Heterocedasticidade

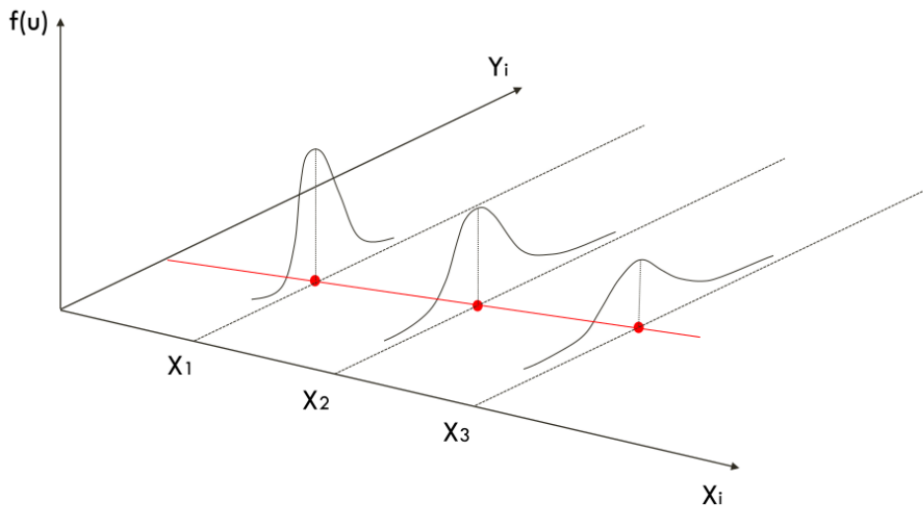
- O parâmetro que controla a dispersão de y_i em torno da média e mede a incerteza do modelo de regressão é a variância.
- Se a dispersão de y_i em torno da média aumenta à medida que x_i aumenta, então a incerteza com relação a y_i cresce com o aumento de x_i , evidenciando que a variância não é constante.
- É necessário procurar uma forma de modelar a variância que aumente com o crescimento de x_i .
- A forma mais comum é considerar:

$$var(y_i) = \sigma_i^2$$

Homocedasticidade



Heterocedasticidade



Algumas fontes de heterocedasticidade

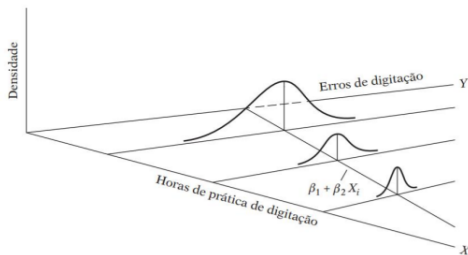
1) Processo de erro e aprendizagem:

O comportamento incorreto pode diminuir ao longo do tempo.

$$y = \beta_2 x_2 + u_i$$

↓ ↘
Erro de digitação Horas de prática

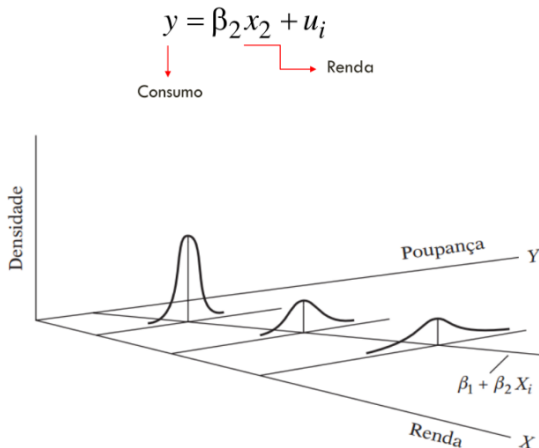
A variância do erro tende a diminuir conforme se aumenta as horas praticadas.



Algumas fontes de heterocedasticidade

2) Renda discricionária

Quando a renda aumenta, a variabilidade do consumo tende a se elevar, assim como sua variância.



Algumas fontes de heterocedasticidade

3) Técnicas de coleta de dados:

Sofisticação dos equipamentos reduz os erros de coleta de dados.

Ex: Bancos que possuem equipamentos mais sofisticados.

$$y = \beta_2 x_2 + u_i$$

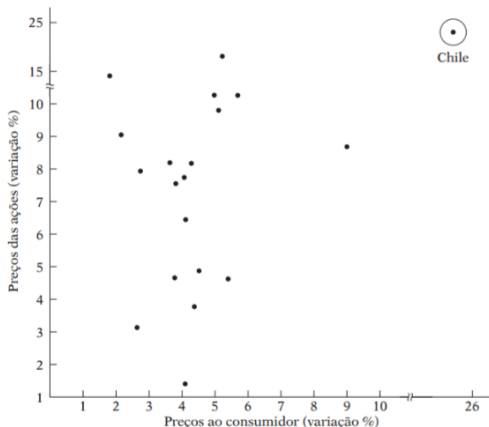
Projeção de receita

Projeção do número de clientes

Algumas fontes de heterocedasticidade

4) Presença de outliers

Outliers são observações discrepantes em relação à média da amostra. A presença de informações discrepantes pode alterar a variância do erro em determinado ponto da amostra.



Algumas fontes de heterocedasticidade

5) Má especificação do modelo:

Variáveis importantes podem estar contidas no termo de erro quando não especificadas.

$$\text{consumo} = \beta_2 \text{riqueza} + u_i \rightarrow \text{Se a variável "renda" estiver contida no termo de erro, a variância de } u_i \text{ será heterocedástica.}$$

6) Assimetria dos regressores

A distribuição dos regressores pode ocasionar heterocedasticidade (desigualdade na renda ou educação, por exemplo).

7) Transformação dos dados:

- Transformação proporcional: pode realçar a dispersão dos dados.
- Transformação em primeira diferença: pode criar clusters de volatilidade.

O que ocorre com as estimativas de MQO na presença de Heterocedasticidade

- Para estabelecer a não tendenciosidade e consistência dos estimadores de mínimos quadrados não é necessário que os termos de erro sejam homocedásticos.
- A heterocedasticidade não provoca viés ou inconsistência nos estimadores de mínimos quadrados.
- Sob certas condições de regularidade, $\hat{\beta}$ ainda é assintoticamente normalmente distribuído.
- A interpretação das medidas de qualidade de ajuste R^2 e R^2 ajustado não é afetada pela presença de heterocedasticidade.

O que ocorre com as estimativas de MQO na presença de Heterocedasticidade

- **PERGUNTA:** Admitindo-se que os estimadores sejam lineares, não tendenciosos e consistentes, eles ainda são eficientes? Isto é, têm variância mínima na classe dos estimadores não tendenciosos?

Os estimador de mínimos quadrados não é mais o melhor estimador linear não tendencioso (BLUE).

- O teorema de Gauss-Markov que diz que os estimadores de MQO são os melhores estimadores lineares não viesados, vale-se da hipótese de homocedasticidade.

O que ocorre com as variâncias dos estimadores de MQO na presença de Heterocedasticidade

O que ocorre com as variâncias dos estimadores de MQO na presença de Heterocedasticidade

O que ocorre com as variâncias dos estimadores de MQO na presença de Heterocedasticidade

O que ocorre com as estimativas de MQO na presença de Heterocedasticidade

- Os erros padrão dos estimadores de MQO não são mais válidos para construirmos intervalos de confiança e testes de hipóteses.
- As estatísticas t habituais dos estimadores MQO não têm distribuições t na presença de heterocedasticidade.
- De modo semelhante, as estatísticas F não têm distribuição F e a estatística da RV não tem distribuição qui-quadrada assintótica.
- E esse problema não é resolvido com o uso de grandes amostras.



Consequências de usar MQO na presença de heterocedasticidade.

I. Estimação de MQO admitindo heterocedasticidade:

- Suponha que usemos o estimador de MQO e a variância considerando heterocedasticidade.
- Usando essa variância e supondo σ_i^2 conhecido ainda não é possível estabelecer intervalos de confiança e testar hipóteses com os testes habituais.
- Uma vez que as variâncias serão desnecessariamente maiores, os testes t e F , provavelmente nos darão resultados imprecisos.
- Por exemplo, se a variância $var(\hat{\beta}_1)$ for excessivamente grande, o que parece ser um coeficiente estatisticamente insignificante (porque o valor de t é menor que o adequado), pode ser significativo.

Consequências de usar MQO na presença de heterocedasticidade.

II. Estimação de MQO desconsiderando a heterocedasticidade:

- Usar a fórmula da variância considerando homocedasticidade é o caso mais provável.
- A variância sob homocedasticidade é um estimador tendencioso da verdadeira variância na presença de heterocedasticidade.
- Em geral, não podemos dizer se o viés é positivo ou negativo pois porque depende da natureza da relação entre σ_i^2 e os valores assumidos pela variável explicativa.
- Não podemos, mais uma vez, considerar os intervalos de confiança e testes de hipóteses usuais.

Quais as possibilidades para resolver esse problema???

- I. Usar uma transformação para estabilizar a variância.**
- II. Uma outra possibilidade de não abandonar de vez o método MQO é obter erro padrão, estatísticas t e F válidos na presença de heterocedasticidade de forma desconhecida. Esse método consiste em estimar corretamente as variâncias dos estimadores na presença de heterocedasticidade.**
- III. Outra possibilidade é usar o Método de mínimos quadrados ponderados que considera a variabilidade desigual da variável dependente e é capaz de produzir estimadores BLUE. Contudo, é necessário especificar a forma da variação.**

Exemplo de uso de transformação para estabilizar a variância

Diversidade de espécies nas Ilhas Galápagos

A relação entre o número de espécies de plantas e várias variáveis foi avaliada para 30 ilhas Galápagos.

- Species: O número de espécies vegetais encontradas na ilha.
- Area: A área da ilha.
- Elevation: A maior elevação da ilha.
- Scruz: A distância da ilha de Santa Cruz.
- Adjacent: A área da ilha adjacente.

Faça um gráfico para verificar a suposição de variância constante. Aplique a transformação: raiz quadrada na variável dependente e repita o gráfico. O que você pode concluir?

Exemplo de uso de transformação para estabilizar a variância

No R



```
require(faraway)
library(stats)
data(gala)
fitgala=lm(Species~Area+Elevation+Scruz+Nearest+Adjacent,
gala)
fitgalat=lm(sqrt(Species)~Area+Elevation+Scruz+Nearest+
Adjacent, gala)

par(mfrow=c (1, 2))
plot(fitgala$fitted.values,residuals(fitgala),xlab="Fitted",
ylab="Residuals",main="Usual")
plot(fitgalat$fitted.values,residuals(fitgalat),xlab="Fitted",
ylab="Residuals",main="Transformed")
```

Detectando a heterocedasticidade

MÉTODOS INFORMAIS

I. Natureza do Problema

- A natureza do problema já pode nos dar algum indício da presença de heterocedasticidade.
- O problema de estimar despesa com alimentação a partir da renda.
- Em uma análise que envolve despesas com investimento em relação a vendas e taxa de juros, em geral espera-se encontrar heterocedasticidade se empresas de tamanho pequeno, médio e grande fizerem parte da amostra.
- Ou seja, em dados de corte transversal, a heterocedasticidade pode ser a regra e não a exceção.

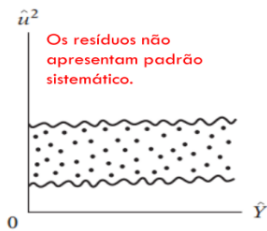
Detectando a heterocedasticidade

MÉTODOS INFORMAIS

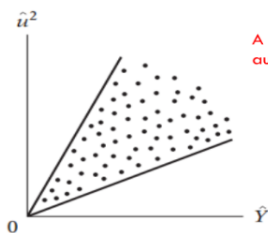
II. Técnicas Gráficas

- É possível fazer a análise de regressão supondo-se que não há heterocedasticidade e fazer o gráfico dos resíduos de mínimos quadrados.
- Se os erros são heterocedásticos o gráfico pode exibir um padrão sistemático.
- Em caso de homocedasticidade, não deve haver padrão de qualquer tipo no resíduo.
- Um gráfico dos resíduos versus valores ajustados ou um gráfico dos resíduos versus variável explicativa dão ideia da existência de algum padrão sistemático.

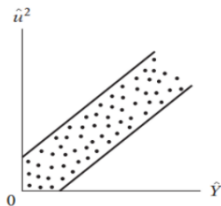
Gráfico dos Resíduos



(a)

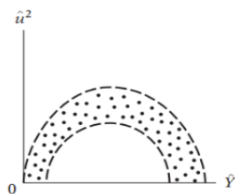


(b)



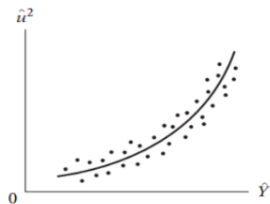
(c)

Relação linear



(d)

Relação quadrática



(e)

Relação quadrática

Gráfico dos Resíduos

- Na Figura (a) vemos que não há padrão sistemático, o que sugere não haver heterocedasticidade nos dados.
- As Figuras (b) a (e), mostram padrões definidos
- As Figuras (b) e (c) sugerem uma relação linear.
- As Figuras (d) e (e) sugerem uma relação quadrática.
- Usando as informações retiradas dos gráficos é possível transformar os dados de modo que na regressão com os dados transformados, a variância do termo de erro seja homocedástica.

Exemplo

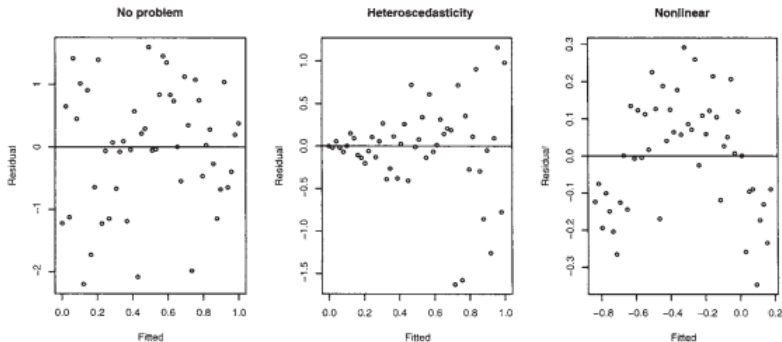


Figure 4.1 *Residuals vs. fitted plots—the first suggests no change to the current model while the second shows nonconstant variance and the third indicates some nonlinearity, which should prompt some change in the structural form of the model.*

Exemplo

No R

Os comandos a seguir geram gráficos que exibem: i) Variância Constante, ii) Variância não-constante (leve), iii) Variância não-constante (forte) e iv) Não-linearidade.

```
par(mfrow=c (2, 2))  
plot (1:50, rnorm (50))  
plot (1:50, sqrt ((1:50))*rnorm(50))  
plot (1:50, (1:50)*rnorm(50))  
plot(1:50, cos ((1:50)*pi/25)+rnorm(50))
```

Detectando a heterocedasticidade

MÉTODOS FORMAIS

I. Teste de Park

- Park sugere que σ_i^2 seja uma função da V.A. X_i a partir da seguinte forma funcional:

$$\ln \sigma_i^2 = \ln \sigma^2 + \beta \ln X_i + \nu_i,$$

em que ν_i é o termo de erro estocástico.

- Uma vez que σ_i^2 , em geral, não é conhecido, Park sugere fazer a regressão de MQO desconsiderando heterocedasticidade e em seguida calcular a regressão:

$$\ln \hat{u}_i^2 = \ln \sigma^2 + \beta \ln X_i + \nu_i,$$

- Se o β for significativo, sugere-se que a heterocedasticidade está presente nos dados.

Ilustrando a abordagem de Park...

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

em que Y = remuneração média em milhares de dólares, X = produtividade média em milhares de dólares e i = i -ésimo tamanho do emprego de estabelecimento. Os resultados da regressão são os seguintes:

$$\begin{aligned}\hat{Y}_i &= 1992,3452 + 0,2329X_i \\ \text{ep} &= (936,4791) \quad (0,0998) \\ t &= (2,1275) \quad (2,333) \quad R^2 = 0,4375\end{aligned}\tag{11.5.3}$$

Então, calcula-se a regressão dos resíduos obtidos na regressão (11.5.3) contra X_i , como sugerido na Equação (11.5.2), dando os resultados a seguir:

$$\begin{aligned}\widehat{\ln \hat{u}_i^2} &= 35,817 - 2,8099 \ln X_i \\ \text{ep} &= (38,319) \quad (4,216) \\ t &= (0,934) \quad (-0,667) \quad R^2 = 0,0595\end{aligned}\tag{11.5.4}$$

Obviamente, não há relação estatisticamente significativa entre as duas variáveis. Seguindo o teste de Park, pode-se concluir que não há heterocedasticidade na variância dos erros.¹³

Detectando a heterocedasticidade

II. Teste de Goldfeld e Quandt

- Goldfeld e Quandt alegam que o termo de erro ν_i pode não satisfazer as suposições de MQO e ele mesmo pode ser heterocedástico.
- Esse teste é aplicável quando se supõe que a variância heterocedástica relaciona-se positivamente com uma das variáveis explicativas, como por exemplo: $\sigma_i^2 = \sigma^2 X_i^2$.
- Ordene as observações de acordo com os valores de X_i .
- Omita c observações centrais e divida as observações remanescentes em dois grupos com $(n - c)/2$ observações em cada um.

II. Teste de Goldfeld e Quandt

- Ajuste as regressões de MQO separadas para os dois grupos e obtenha as respectivas somas dos quadrados de resíduos.
- Calcule $GQ = \frac{SQR_2/gl_2}{SQR_1/gl_1}$, em que SQR_1 representa a SQR a partir da regressão correspondente aos valores menores de X_i (grupo de pequena variância) e SQR_2 a partir do conjunto com maiores valores de X_i (grupo com variância maior), $gl_1 = gl_2 = \frac{(n-c)}{2} - k$.
- Rejeite a hipótese nula de variâncias iguais se $GQ > F_c$.

II. Teste de Goldfeld e Quandt

- F_c é o valor crítico da distribuição F ao nível de significância escolhido e k o número de parâmetros estimados no modelo.
- É importante ressaltar que as c observações são omitidas para acentuar a diferença entre o grupo com variâncias pequenas (SQR_1) e o de grandes variâncias (SQR_2).
- Goldfeld-Quandt sugerem $c = 8$ se n for aproximadamente 30 e $c = 16$ se n for aproximadamente 60.

Ilustrando a abordagem de Goldfeld-Quandt...

		Dados ordenados por valores de X	
consumo	renda	Y	X
Y	X		
55	80	55	80
65	100	70	85
70	85	75	90
80	110	65	100
79	120	74	105
84	115	80	110
98	130	84	115
95	140	79	120
90	125	90	125
75	90	98	130
74	105	95	140
110	160	108	145
113	150	113	150
125	165	110	160
108	145	125	165
115	180	115	180
140	225	130	185
120	200	135	190
145	240	120	200
130	185	140	205
152	220	144	210
144	210	152	220
175	245	140	225
180	260	137	230
135	190	145	240
140	205	175	245
178	265	189	250
191	270	180	260
137	230	178	265
189	250	191	270

- I. Não existe regra para o tamanho de c (observações excluídas no teste).
- II. Podemos adotar $c = 4$, para $n = 30$.
- III. Estimamos 2 regressões com 13 observações.

4 observações do meio

Ilustrando a abordagem de Goldfeld-Quandt...

Regressão baseada nas 13 primeiras observações:

$$\hat{Y}_i = 3,4094 + 0,6968X_i$$

(8,7049) (0,0744) $r^2 = 0,8887$ $SQR_1 = 377,17$ $gl = 11$

Regressão baseada nas 13 últimas observadas:

$$\hat{Y}_i = -28,0272 + 0,7941X_i$$

(30,6421) (0,1319) $r^2 = 0,7681$ $SQR_2 = 1536,8$ $gl = 11$

Desses resultados, obtemos

$$\lambda = \frac{SQR_2/gl}{SQR_1/gl} = \frac{1536,8/gl}{377,17/gl}$$
$$\lambda = 4,07$$

O valor crítico de F para 11 graus de liberdade no numerador e no denominador no nível de 5% é 2,82. Uma vez que o $F (= \lambda)$ estimado excede o valor crítico, podemos concluir que há heterocedasticidade na variância de erro. Entretanto, se o nível de significância for fixado em 1%, não podemos rejeitar a suposição de homocedasticidade. (Por quê?) Note que o p valor do λ observado é 0,014.

III. Teste de Breusch-Pagan

- O sucesso do teste de Goldfeld-Quandt depende do valor de c e da identificação da variável X correta com a qual se colocam as observações em ordem.
- Considere um modelo de regressão com k variáveis explicativas.
- Suponha que $\sigma_j^2 = \alpha_1 + \alpha_2 Z_{2j} + \dots + \alpha_m Z_{mj}$.
- Alguns ou todos os X podem servir como Z .
- Testar a hipótese de homocedasticidade equivale a testar se $\alpha_2 = \dots = \alpha_m = 0$.
- Obtenha os resíduos via regressão de MQO.

Detectando a heterocedasticidade

III. Teste de Breusch-Pagan

- Obtenha $\tilde{\sigma}^2 = \sum \hat{u}_i^2 / n$ (EMV de σ^2).
- Construa $p_i = \hat{u}_i^2 / \tilde{\sigma}^2$.
- Faça a regressão $p_i = \alpha_1 + \alpha_2 Z_{2i} + \dots + \alpha_m Z_{mi} + \nu_i$, em que ν_i é o termo residual.
- Obtenha SQE da regressão anterior e defina $BP = SQE/2$.
- Supondo que os u_i sejam normalmente distribuídos e sob homocedasticidade temos que BP segue distribuição qui-quadrado com $(m - 1)$ gl.
- Rejeite a hipótese nula de homocedasticidade se BP for maior que o valor crítico da Qui-quadrado no nível de significância escolhido.

Ilustrando a abordagem de Breusch-Pagan...

Etapa 1.

$$\hat{Y}_i = 9,2903 + 0,6378X_i$$
$$\text{ep} = (5,2314) \quad (0,0286) \quad \text{SQR} = 2361,153 \quad R^2 = 0,9466 \quad (11.5.18)$$

Etapa 2.

$$\hat{\sigma}^2 = \sum \hat{u}_i^2 / 30 = 2361,153 / 30 = 78,7051$$

Etapa 3. Divida os resíduos elevados ao quadrado \hat{u}_i obtidos da regressão (11.5.18) por 78,7051 para construir a variável p_i .

Etapa 4. Supondo que os p_i sejam linearmente relacionados a $X_i (= Z_i)$ como na Equação (11.5.14), obtemos a regressão

$$\hat{p}_i = -0,7426 + 0,0101X_i$$
$$\text{ep} = (0,7529) \quad (0,0041) \quad \text{SQE} = 10,4280 \quad R^2 = 0,18 \quad (11.5.19)$$

Etapa 5.

$$\Theta = \frac{1}{2}(\text{SQE}) = 5,2140$$

qui-quadrado a 5% é 3,8414
(11.5.20)

Detectando a heterocedasticidade

IV. Teste de White

- O teste de Breusch-Pagan é sensível à hipótese de normalidade.
- O teste geral de heterocedasticidade proposto por White não requer a hipótese de normalidade.

- Considere o modelo de regressão

$$y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i.$$

- Obtenha os resíduos \hat{u}_i da regressão anterior e faça a seguinte regressão auxiliar:

$$\hat{u}_i^2 = \alpha_1 + \alpha_2 X_{2i} + \alpha_3 X_{3i} + \alpha_4 X_{2i}^2 + \alpha_5 X_{3i}^2 + \alpha_6 X_{2i} X_{3i} + \nu_i$$

- Obtenha o R^2 dessa regressão auxiliar.

Detectando a heterocedasticidade

IV. Teste de White

- Sob a hipótese nula de que não há heterocedasticidade

$$WT = nR^2 \sim \chi_{gl}^2,$$

em que gl é o número de regressores na regressão auxiliar.

- Rejeite a hipótese nula de homocedasticidade ($\alpha_2 = \alpha_3 = \dots = \alpha_k = 0$) se WT for maior que o valor crítico da Qui-quadrado no nível de significância escolhido.
- Uma desvantagem desse teste é o número alto de graus de liberdade devido ao número de regressores introduzidos.

Ilustrando a abordagem de White...

$$\begin{aligned}\widehat{u}_i^2 = & -5,8417 + 2,5629 \ln \text{Trade}_i + 0,6918 \ln \text{PIB}_i \\ & - 0,4081(\ln \text{Trade}_i)^2 - 0,0491(\ln \text{PIB}_i)^2 \\ & + 0,0015(\ln \text{Trade}_i)(\ln \text{PIB}_i)\end{aligned}\quad R^2 = 0,1148 \quad (11.5.25)$$

Nota: os erros padrão não são apresentados, pois não são pertinentes para nossos fins.

Agora $n \cdot R^2 = 41 \ (0,1148) = 4,7068$ tem, assintoticamente, uma distribuição quiquadrado com 5 graus de liberdade (por quê?). O valor crítico de 5% graus de liberdade para o qui-quadrado e nível de significância de 5% é de 11,0705 e com significância de 10% é de 9,2363 e com 25% é de 6,62568. Para fins práticos, podemos concluir, com base no teste de White, que não há heterocedasticidade.

Detectando a heterocedasticidade

V. Teste de Koenker

- É bastante utilizado devido a sua simplicidade e também não requer a hipótese de normalidade.
- Nesse teste os resíduos ao quadrado são regredidos contra os valores estimados do regressando elevados ao quadrado.
- Considere o modelo de regressão
$$y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_k X_{ki} + u_i.$$
- Faça a seguinte regressão auxiliar:

$$\hat{u}_i^2 = \alpha_1 + \alpha_2 (\hat{y}_i)^2 + \nu_i$$

- A hipótese nula de que $\alpha_2 = 0$, pode ser testada pelo teste usual t e F considerando k graus de liberdade.

Observações sobre os testes de Heterocedasticidade

- Alguns testes se baseiam nas suposições do modelo de regressão. Caso alguma hipótese seja violada, um teste pode rejeitar H_0 mesmo quando a variância seja constante.
- Alguns autores recomendam aplicar um teste de especificação correta antes de aplicar um teste de heterocedasticidade.



Detectando a heterocedasticidade

EXERCÍCIO 1: Considere os dados contidos no arquivo "hprice1(wooldridge)" do R.

- a) Encontre um modelo para explicar o preço dos imóveis.
- b) Faça o gráfico dos resíduos e verifique a suposição de homocedasticidade.
- c) Aplique os testes do R para verificar a suposição de homocedasticidade.
- d) É possível fazer um teste de igualdade de variâncias? Como seria?
- e) Coloque as variáveis do modelo em forma logarítmica. O que ocorre com a heterocedasticidade?
- f) Interprete os coeficientes.

Exercício 1

Detectando a heterocedasticidade

EXERCÍCIO 2

- a) Aplique as técnicas gráficas e testes de hipóteses no R para os seguintes dados:
- Dados "PublicSchools" do R.
 - Dados "attitude" do R.
 - DadosReg (Consumo combustível por hab)
 - Renda Paraíba
- b) Quais testes são mais indicados em cada situação?
- c) Existe implementado no R outro teste para detectar a heterocedasticidade?
- d) Aplique os testes aos mesmos conjuntos de dados do item anterior.

Exercício 1